The distribution of longest run lengths in integer compositions

Herbert S. Wilf

June 29, 2009

Abstract

We find the generating function for $C(n,k,r)$, the number of compositions of n into k positive parts all of whose runs (contiguous blocks of constant parts) have lengths less than r, using recent generalizations of the method of Guibas and Odlyzko for finding the number of words that avoid a given list of subwords.

1 Introduction

A composition of an integer n is a representation $n = a_1 + a_2 + \cdots + a_k$ in which the parts a_i are positive integers, and where the order of the parts is important. Thus $9 = 1+1+1+4+2$ is one of the compositions of $n = 9$, and $9 = 4+1+1+2+1$ is another.

A *run* in a composition is a maximal string of consecutive identical parts. The composition

$$28 = 3 + 5 + 5 + 5 + 3 + 3 + 4$$

has run lengths of 1,3,2,1, for example. In this note we find (the generating function of) $C(n,k,r)$, the number of compositions of n into k parts, whose runs all have lengths $< r$ (see Theorem 3 below), by using recent generalizations of the Guibas-Odlyzko theory of counting words that avoid a given list of subwords.

In their 1981 paper [4], Guibas and Odlyzko gave an elegant solution to the following counting problem. Given an alphabet $\mathcal{A}$, and a list $\mathcal{L}$ of words over that alphabet, the list being *reduced* in the sense that no word on the list is a subword of any other. How many words of length n do not contain any of the words in $\mathcal{L}$ as a subword? Other solutions of this problem have been given by the cluster method of Goulden and Jackson [2], and by Zeilberger's [10] method of counting words that avoid "mistakes."

The results of [4] have recently been extended by A.N. Myers [9] to the situation wherein the letters of the alphabet are assigned weights, the weight of a word is the sum of the weights of its letters, and one is to find the number of words of weight n that avoid the members of the list $\mathcal{L}$. This allows us to solve problems involving compositions of integers as well as problems that do not involve compositions.

Finally, Myers's results have been complemented by Heubach and Kitaev [5] to provide the number of words of *length k and weight n* that avoid the members of the list $\mathcal{L}$, though their theorems are restricted to the alphabet $\{1, 2, \ldots, n\}$ and therefore apply almost exclusively to integer compositions.

The above theorems present the generating function for the desired numbers of words as the first component of the solution vector of a system of linear, simultaneous equations, or, by using Cramer's rule, as a ratio of two determinants.

The main point of this note is the following. The easy case in such word problems is the case in which every pair of distinct words on the forbidden list $\mathcal{L}$ has correlation 0, in a sense to be explained below, or equivalently, for every pair x, y of distinct words on that list, no suffix of x is also a prefix of y. In that situation, the matrix of coefficients of the system of linear equations that expresses the answer to the question has a very simple form. It consists of a nonzero first row and first column and main diagonal, all other entries being 0's.

For a matrix of that form it is easy to write out the solution of the governing system of linear equations simply and explicitly. We will do that below and then find the generating function for $C(n, k, j)$, the number of compositions of n into k parts the lengths of whose runs is at most j.

2 The main theorem

Let X and Y be two words over a given alphabet. We define *the correlation c_{XY} of X on Y*, as follows.

- Write the word X above the word Y, aligned so that the rightmost letter of X is above the rightmost letter of Y.
- Fix some integer $j \geq 0$. Shift Y j places to the left, so the rightmost letter of Y is now under the $(j + 1)$st letter of X, counting from the right.
- Examine the subword of X that now overlaps with Y. This is the maximal prefix of X that has letters of the shifted Y below it.

- If that subword of X is identical with the subword of Y that lies below it, take $c_j = 1$, else take $c_j = 0$.
- Having done this for all j, the correlation of X on Y is the binary vector $c_0c_1c_2\ldots$.

For example, if $X = 110$ and $Y = 1011$ then $c_{XY} = 011$ and $c_{YX} = 0010$, in which we have written the bits of the c's in the order $c_0c_1\ldots c_{m-1}$.

Let each letter u of the alphabet be assigned a weight $w(u)$, and let the weight of a word be the sum of the weights of its letters. Finally, if X is an m-letter word $X = a_0a_1\ldots a_{m-1}$, define the *correlation polynomial* $c_{XY}(x, q)$ of X on Y to be

$$c_{XY}(x,q) = c_0 + c_1x^{w(a_{m-1})}q + c_2x^{w(a_{m-2}a_{m-1})}q^2 + \cdots + c_{m-1}x^{w(a_1a_2\ldots a_{m-1})}q^{m-1}. \qquad (1)$$

The main result of [5], which extends the main result of [9], which in turn extends the main result of [4], is the following.

Theorem 1 (Heubach, Kitaev) *Let $\mathcal{L} = \{S_1, \ldots, S_k\}$ be a list of integer compositions, such that no composition on the list is contained in any other. Let $F(x,q) = \sum_\sigma x^{w(\sigma)}q^{\ell(\sigma)}$, the sum being extended over all compositions of all integers that avoid every word on the list $\mathcal{L}$, where $\ell(\sigma)$ is the length of the word (number of parts of) σ and $w(\sigma)$ is the sum of the parts of σ. Then $F(x,q)$ is the component x_1 of the solution vector of the following system of linear equations:*

$$\begin{pmatrix} 1-x(1+q) & 1-x & \ldots & 1-x \\ x^{w(S_1)}q^{\ell(S_1)} & -c_{11}(x,q) & \ldots & -c_{1k}(x,q) \\ \vdots & \vdots & \ddots & \vdots \\ x^{w(S_k)}q^{\ell(S_k)} & -c_{k1}(x,q) & \ldots & -c_{kk}(x,q) \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_{k+1} \end{pmatrix} = \begin{pmatrix} 1-x \\ 0 \\ \vdots \\ 0 \end{pmatrix} \qquad (2)$$

3 The easy case

We now specialize to the case where $c_{ij}(x, q) = 0$ for all $i \neq j$, $1 \leq i, j \leq k$, k being the length of the forbidden word list $\mathcal{L}$. The coefficient matrix entries in the equations (2) then all vanish except for those in the first row, the first column, and the main diagonal. For any such matrix, B, say, the first entry of the solution vector of the equations $B\mathbf{x} = (1 - x, 0, \ldots, 0)^T$ is easily verified to be

$$x_1 = \frac{1-x}{b_{11} - b_{12}\frac{b_{21}}{b_{22}} - \cdots - b_{1,k+1}\frac{b_{k+1,1}}{b_{k+1,k+1}}}.$$

If we apply this result to the equations (2) we obtain

Theorem 2 *Let $\mathcal{L} = \{S_1, \ldots, S_k\}$ be a list of integer compositions, such that no word on the list is contained in any other. Suppose further that all correlation polynomials $c_{ij}(x, q) = 0$, for $i \neq j$. Let $F(x, q) = \sum_\sigma x^{w(\sigma)}q^{\ell(\sigma)}$, the sum being extended over all compositions σ that avoid every word on the list $\mathcal{L}$, where $\ell(\sigma)$ is the length of the word σ. Then we have the explicit formula*

$$F(x,q) = \frac{1}{1 - \frac{qx}{1-x} + \sum_{j=1}^{k} \frac{x^{w(S_j)}q^{\ell(S_j)}}{c_{j,j}(x,q)}}. \tag{3}$$

4 Carlitz compositions and beyond

4.1 Carlitz compositions

We apply the results of the previous section to finding the distribution function of the lengths of the longest runs of integer compositions. Again, a run in a composition is a maximal string of identical parts. The composition 28=3+5+5+5+3+3+4 has run lengths of 1,3,2,1, for example.

A Carlitz composition is one all of whose runs have length 1. That is, a Carlitz composition is one in which no two consecutive parts are equal. These compositions have been extensively studied in recent years, both exactly and asymptotically [1, 7, 8]. The machinery of "the easy case" above counts Carlitz compositions of n, as follows.

The list $\mathcal{L}$ of forbidden subwords is $\mathcal{L} = \{11, 22, 33, 44, \ldots\}$. A Carlitz composition is evidently one that avoids this list, and also evidently, this list belongs to the easy case, i.e., the off-diagonal correlation polynomials all vanish. Thus we can use Theorem 2.

The word S_j is jj, and its weight is $w(S_j) = 2j$. The correlation polynomials $c_{S_iS_j}$ vanish for all $i \neq j$, while for $i = j$ we have by (1),

$$c_{S_jS_j}(x, y) = 1 + x^j q.$$

If $C(n, k)$ is the number of Carlitz compositions of n into k parts, we now have from equation (3),

$$\begin{aligned}\sum_{n,k} C(n,k)x^n q^k &= \frac{1}{1 - \frac{xq}{1-x}) + q^2 \sum_{j \geq 1} \frac{x^{2j}}{1+qx^j}} \\ &= 1 + qx + qx^2 + (q + 2q^2)x^3 + (q + 2q^2 + q^3)x^4 + (q + 4q^2 + 2q^3)x^5 + \ldots\end{aligned}$$

This generating function has previously been found, in somewhat different form, by Knopfmacher and Prodinger [7].

4.2 Beyond

Now we find the distribution function of the maximum run length in compositions of n that have k parts.

Let $C(n, k, r)$ denote the number of compositions of n into k parts that have no run of length $\geq r$. Note that $C(n, k, 2)$ counts Carlitz compositions of n with k parts. To find $C(n, k, r)$ we use the list $\mathcal{L} = \{1^r, 2^r, 3^r, \dots\}$ of forbidden words, where, e.g., 1^r is a string of r 1's. Then again the list $\mathcal{L}$ qualifies for "the easy case," since the correlations all vanish off of the diagonal. while on the diagonal,

$$c_{S_j S_j}(x, y) = 1 + x^j q + x^{2j} q^2 + \cdots + x^{j(r-1)} q^{r-1} = \frac{1 - q^r x^{rj}}{1 - qx^j}.$$

We now have from equation (3),

Theorem 3 *The number $C(n, k, r)$ of compositions of n into k parts that have no run of length $\geq r$ has the generating function*

$$\sum_{n,k} C(n, k, r) x^n q^k = \frac{1}{1 - \frac{xq}{1-x} + q^r \sum_{j \geq 1} \frac{x^{rj}(1-qx^j)}{1-q^r x^{rj}}}. \tag{4}$$

When $r = 3$ we have

$$\sum_{n,k} C(n, k, 3) x^n q^k = 1 + qx + (q + q^2)x^2 + (q + 2q^2)x^3 + (q + 3q^2 + 3q^3)x^4 + \dots,$$

and for $r = 4$,

$$\sum_{n,k} C(n, k, 4) x^n q^k = 1 + qx + (q + q^2)x^2 + (q + 2q^2 + q^3)x^3 + (q + 3q^2 + 3q^3)x^4 + \dots.$$

The average length of the longest run in a composition of n has been found to be $\sim \log_2 n$, by Grabner et al [3], using the method of i.i.d. geometric random variables.

References

[1] L. Carlitz, Restricted compositions, The Fibonacci Quart., **14** (1976), 254264.

[2] I. P. Goulden and D. M. Jackson, An inversion theorem for cluster decompositions of sequences with distinguished subsequences, J. London Math. Soc. **20** (1979), no. 3, 567–576.

[3] P. Grabner, A. Knopfmacher and H. Prodinger, Combinatorics of geometrically distributed random variables: Run statistics, Theoretical Computer Science **297** (2003), 261–270.

[4] L.J. Guibas and A.M. Odlyzko, String overlaps, pattern matching, and nontransitive games, J. Combinatorial Theory, Ser. A, **30** (1981), 183–208.

[5] Silvia Heubach and Sergey Kitaev, Avoiding substrings in compositions, `arXiv:math/0903.5135` [math.CO]

[6] S. Heubach and T. Mansour, Enumeration of 3-letter patterns in compositions, `arXiv:math/0603285v1` [math.CO].

[7] Arnold Knopfmacher and Helmut Prodinger, On Carlitz compositions, European J. Combin. **19** (1998), no. 5, 579–589.

[8] Guy Louchard and Helmut Prodinger, Probabilistic analysis of Carlitz compositions, Discrete Math. Theor. Comput. Sci. **5** (2002), no. 1, 71–95.

[9] Amy N. Myers, Forbidden substrings on weighted alphabets, Australasian J. Math., to appear.

[10] Doron Zeilberger, Enumeration of words by their number of mistakes. Discrete Math. **34** (1981), no. 1, 89–91.